36-303: Sampling, Surveys and Society
Exam 2 Solutions

1. [20 pts] *Multiple Choice (4 parts). For each part, circle the roman numeral of the one best answer.*

   (a) [5 pts] . . . older people were less likely to report their income, and among people who did report incomes, older people tended to earn more than younger people.

   Which imputation method below is **not** recommended (or, circle iv. if all are OK)?

   i. For each missing income, fill in the mean of all of the incomes you do have.

   (b) [5 pts] . . . If we make the groups so that *observations* **within groups** *are more* **similar** *to each other*, and *observations* **between groups** *are more* **different** *from each other*, then, all other things being equal, we expect

   ii. The variance of the stratified sample mean $\overline{y}_{st}$ will go **down** *and the variance of the cluster sample mean $\overline{y}_{cl}$ will go* **up**.

   (c) [5 pts] Which of the following is **not** one of the recommended things to work on, to reduce the tendency of survey subjects to not respond?

   ii. Choice of stratified or cluster sampling.

   (d) [5 pts] In one-stage clustered sampling, the ICC ρ measures

   iii. The correlation between observations in the same cluster.

2. [20 pts] *Cluster Sampling (3 parts).*

   . . . joys of summer. . . One roadside stand sells corn in bunches of 3 ears each; they have 580 such bunches for sale. A food inspector samples 12 of these bunches and counts the number of worm fragments he finds on each ear in all 12 bunches, to determine the average number of worm fragments per ear of corn. The inspector's data are as follows:

	Bunch											
	01	02	03	04	05	06	07	08	09	10	11	12
Ear 1	1	4	0	3	4	0	5	3	7	3	4	0
Ear 2	5	2	1	6	9	7	5	0	3	1	7	0
Ear 3	7	4	2	6	8	3	1	2	5	4	9	0
mean	4.33	3.33	1.00	5.00	7.00	3.33	3.67	1.67	5.00	2.67	6.67	0.00

$$s^2_{\overline{y}_i} = \frac{1}{12-1}\sum_{i=1}^{12}(\overline{y}_i - \overline{y}_{cl})^2 = 4.53$$

$$s^2_{y_{ij}} = \frac{1}{36-1}\sum_{i=1}^{12}\sum_{j=1}^{3}(y_{ij} - \overline{y}_{srs})^2 = 7.38$$

   (a) [6 pts] Compute $\overline{y}_{cl}$ and $\overline{y}_{srs}$; explain why they should be same/different.

   One-stage cluster sampling with equal cluster sizes is self-weighting, *which means that $\overline{y}_{cl} = \overline{y}_{srs}$. It is easiest to calculate*

$$\overline{y}_{cl} = \frac{1}{12}(4.33 + 3.33 + 1.00 + 5.00 + 7.00 + 3.33 + 3.67 + 1.67 + 5.00 + 2.67 + 6.67 + 0.00) = 3.64$$

(b) [6 pts] Compute SE($\overline{y}_{cl}$) and use this to create an approximate 95% confidence interval for the mean number of worm fragments per ear in the entire population of $580 \times 3 = 1740$ ears of corn.

$$\text{Var}(\overline{y}_{cl}) = (1 - n/N)\frac{1}{n}s^2_{\overline{y}_i} = (1 - 12/580)\frac{1}{12}4.53 = 0.3697$$

so

$$SE(\overline{y}_{cl}) = \sqrt{0.3697} = 0.61$$

The confidence interval can be either of the following (full credit for either one):

$$3.64 \pm (1.96)(0.61) = (2.44, 4.84)$$

or

$$3.64 \pm (2)(0.61) = (2.42, 4.86)$$

(c) [8 pts] Estimate

- The DEFF (design effect) for this design; and
- The ICC (intra-cluster correlation, ρ).

$$DEFF = \frac{Ms^2_{\overline{y}_i}}{s^2_{y_{ij}}} = \frac{(3)(4.53)}{7.38} = 1.84$$

so

$$\rho = (DEFF - 1)/(M - 1) = (1.84 - 1)/(3 - 1) = 0.42$$

3. [18 pts] *Yale Work & Life Survey, Part I (3 parts).*

…there were 2,707 men and 2,609 women enrolled at Yale in 2005–2006, the academic year in which Brescoll did her study.

In Brescoll's survey, 154 men responded, of which 134 planned to become parents someday, and 315 women responded, of which 247 planned to be parents someday[1].

(a) [6 pts] Treating the sample of men as an SRS w/o replacement, compute $\hat{p}_{men}$, the proportion of men at Yale who plan on becoming parents, and its standard error, SE($\hat{p}_{men}$).

$$\hat{p}_{men} = 134/154 = 0.87$$

so

$$\text{Var}(\hat{p}_{men}) = (1 - 154/2707)\frac{1}{154}(0.87)(1 - 0.87) = 0.0006926$$

and so

$$\text{SE}(\hat{p}_{men}) = \sqrt{0.0006926} = 0.026$$

[1] Other Likert-type questions more directly addressed the tradeoff between career and family, but we will just work with this yes/no question.

(b) [6 pts] Do the same for women: compute $\hat{p}_{women}$ and SE($\hat{p}_{women}$).

$$\hat{p}_{women} = 247/315 = 0.78$$

so

$$\text{Var}\,(\hat{p}_{women}) = (1 - 315/2609)\frac{1}{315}(0.78)(1 - 0.78) = 0.0004790$$

and so

$$\text{SE}(\hat{p}_{women}) = \sqrt{0.0004790} = 0.022$$

(c) [6 pts] Test whether there is a significant difference between men and women on this question. . .

$$z = \frac{\hat{p}_{men} - \hat{p}_{women}}{\sqrt{\text{SE}(\hat{p}_{men})^2 + \text{SE}(\hat{p}_{women})^2}} = \frac{0.87 - 0.78}{\sqrt{0.026^2 + 0.022^2}} = 2.63\,;$$

this value is well beyond ± 1.96 *(limits for a Normal 95% confidence interval) and so we reject the null hypothesis that* $p_{men} - p_{women} = 0$ *at lower than the* 0.05 *level, using a 2-sided z-test comparing two sample proportions.*

The conclusion we draw is that a significantly greater proportion of men than women at Yale plan to be parents someday.

(Depending on how much of the burden of parenthood men expect their wives to carry, this may at least partly contradict the earlier finding in the New York Times article.)

4. [24 pts] *Yale Work & Life Survey, Part II (4 parts).*

(a) [6 pts] Combining men and women, and treating the survey as an SRS w/o replacement, estimate the overall proportion of students at Yale who plan on becoming parents. (You do not have to compute the SE).

$$\hat{p}_{srs} = \frac{134 + 247}{154 + 315} = 0.8124$$

(b) [6 pts] We really do not know if Brescoll's survey was an SRS, what the response rate was, etc., but it seems clear there is an imbalance among the respondents between men & women, vs. the Yale student population. Using sex to define two post-strata, compute post stratification weights to apply to this survey.

$$w_{men} = (2707/(2707 + 2609))/(154/(154 + 315)) = 1.55$$

and

$$w_{women} = (2609/(2707 + 2609))/(315/(154 + 315)) = 0.73$$

(c) [6 pts] Using the weights you computed in part (b), compute the weighted proportion of students at Yale who plan on becoming parents (you do not have to compute the SE).

$$\hat{p}_w = \frac{w_{men}(yesses)_{men} + w_{women}(yesses)_{women}}{w_{men}(total)_{men} + w_{women}(total)_{women}} = \frac{(1.55)(134) + (0.73)(247)}{(1.55)(154) + (0.73)(315)} = 0.8279$$

(d) [6 pts] Use the values of the weights to explain the difference between your answers in parts (a) and (c).

$\hat{p}_w$ *is bigger than* $\hat{p}_{srs}$ *because the men, who had the higher proportion thinking they would be parents someday, are weighted more heavily (1.55 per person) than the women (0.73 per person).*

5. [18 pts] A 36-303 group wants to do a survey of student attitudes toward Carnegie Mellon sports teams and sporting events. In their survey proposal they specify the *target population* to be all currently enrolled undergraduates at Cargnegie Mellon, and they indicate that to collect data they will advertise for volunteers to take the survey on FaceBook.com, with a link to a formal survey instrument at QuestionPro.com.

   (a) [6 pts] Did this group specify a *sampling frame* for their survey? If so, say what it is. If not, specify a sampling frame that the group could work with.

   Here are two, different, acceptable answers:

   - *They didn't explicitly state a sampling frame, but the sampling frame is effectively CMU students with facebook accounts.*
   - *They didn't specify a sampling frame. An acceptable, usable sampling frame would be the list of undergraduates in C-Book.*

   (b) [6 pts] Write down **two** very likely sources of *coverage error* for this survey as currently proposed.

   Here are two sources of coverage error; other answers may also be acceptable.

   - *Not all CMU undergraduates have facebook accounts*
   - *Since it is a volunteer survey, only volunteers will respond. As we know, volunteers are often different from the target population at large (for example, they care enough to volunteer!).*

   (c) [6 pts] Make one big suggestion that this group could implement to improve their sampling plan.

   Here is one acceptable answer; other answers may also be acceptable:

   Using C-Book as a sampling frame, take an SRS without replacement from C-Book. Contact those students and invite them to take the survey on QuestionPro.com (using email, FaceBook, or any other way that gives you a high chance of actually getting to them).